SpA-Former:An Effective and lightweight Transformer for image shadow removal

Xiaofeng Zhang[1], Yudi Zhao[1], Chaochen Gu[1]*,
Changsheng Lu[2], Shanying Zhu[1]
[1] *Department of Automation, Shanghai Jiao Tong University, Shanghai, China*
[2] *College of Engineering and Computer Science, The Australian National University, Canberra, Australia*
{framebreak, yudizhao, jacygu, shyzhu}@sjtu.edu.cn,
changshengluu@gmail.com

***Abstract*—In this paper, we propose an Effective and lightweight Transformer for image shadow detection and removal named SpA-Former to recover a shadow-free image from a single shaded image. In contrast to conventional methods that require two stages for shadow detection and then shadow removal, the SpA-Former is a one-stage network capable of learning the mapping function between shadows and no shadows, and does not require a separate shadow detection. SpA-Former is composed of Transformer encoder and CNN decoder, where the CNN decoder contains the GAN network. In the Transformer encoding stage, Gated Feed-Forward Network(GFFN) is devised to control the information flow. In the CNN decoding stage, Two-wheel RNN joint spatial attention(TWRNN) and Fourier transform residual block (FTR) are designed to achieve satisfactory results in shadow removal. The combination of Transformer and CNN is able to feed global features from the Vision Transformer encoder into CNN to enhance the global perception of CNN branches, taking into account the complementarity of local features and the global. The SpA-Former's inference speed is 0.0459s, and the final Parameters and FLOPS are only 0.47MB and 15G, achieving the current lightweight of image shadow removal. The source code of MemoryNet can be obtained from https://github.com/zhangbaijin/SpA-Former-shadow-removal**

***Index Terms*—Shadow removal, light weight, Fourier transform residual block, Gated Feed-Forward Network**

I. Introduction

The existence of shadows in natural images can provide information about the environment and illumination conditions, which in turn us to interpret the situation in the image, but also makes image processing more complex. Therefore shadows are often eliminated in the first stage of image processing, and the critical factor affecting the quality of elimination is the detection and localization of them. Compared to the traditional image processing techniques [41], [42], neural network [37], [38] has powerful representation ability. Recently, there are some works on deep-learning based shadow removal, such as self-supervised models based on GAN [1]–[8], semi-supervised models with mask-assisted guidance [9], [10], [35], [40], and some unsupervised learning methods [11], [16], [18]–[22].

Currently, there are three major issues for shadow removal. First, there is much information hidden by the shadows, it is challenging to remove shadow while preserving the initial image details. Second, the present network is incapable of accounting for the spatial correlation between remote area. The third and most important point, current methods for image de-shadowing are almost always two-stage (detecting shadows first and then removing them), with a large number of parameters, while running very slowly and not being able to be deployed to the mobile side of the robot.

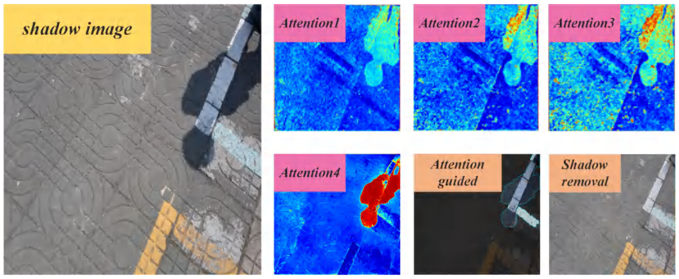

Fig. 1. The process of SpA-Former's shadow removal

To address these problems, we propose An Effective and lightweight Transformer for image shadow detection and removal named SpA-Former. It is difficult to obtain shadow occlusion, since the network uses all functional nodes in the decoder for capturing the global dependencies, but ignores local information. Therefore, considering the core spatial domain information in the shadow image, we design a Two-wheel RNN with the spatial attention mechanism. Spatial attention can use the spatial domain information in the image as a corresponding spatial transformation so that key information can be extracted. On the other hand, in most image de-shadowing architectures, employing multiple layers of ResBlock is a common practice to learn the difference between shadow and non-shadow image pair. Reconstructing a non-shadow image from its shadow counterpart requires filtering of both low and high frequency components. However, traditional ResBlock is good at capturing the high-frequency components of an image, but tends to ignore the low-frequency ones. Furthermore, ResBlock often fails to accurately model the long-range information, thus reducing the effectiveness of de-shadowing. In this paper, we introduce the Fourier transform residual module for integrating low-frequency and high-frequency residual information while capturing both long-term

* Corresponding author: Chaochen Gu.

and short-term interactions.

Last but not least, almost methods of shadow removal are stacks of algorithms that can do well in terms of metrics but are not able to be deployed in real time, so we designed a lightweight network based on Transformer's encoder, decoded as a CNN, while in Transformer's encoding network the GFFN module (Gated Feed-Forward Network) is designed to allows each level to focus on details that are complementary to other levels, we reduce the expansion ratio to reduce the parameters and computational burden, and finally the Params and Flops of SpA-Former only use 0.47MB and 15G, which achieves the parameter minimization of the current algorithm. In summary, our contributions are as follows:

1) This paper introduces a effective and lightweight Transformer for image shadow detection and removal named SpA-Former, which is the first one-stage and lightweight network for shadow removal, in both params and flops, this paper achieves SOTA.
2) This paper proposes a two-wheel joint spatial attention(TWRNN) and Fourier transform residual block(FTR), converting the spatial domain information into an image, while considering the relationship of remote dependencies.
3) The experimental results on ISTD and SRD datasets prove that the proposed method achieves better performance and proves the effectiveness of the network.

II. Realted work

A. Image shadow removal

The earliest weakly supervised first for DeshadowNet [1] has the biggest feature of the fully automatic end-to-end implementation of shadow removal, while the biggest contribution of this paper is to propose a dataset SRD (A New Dataset for Shadow Removal). The authors of the second ST-CGAN [2] propose a multitasking perspective that differs from all existing approaches in that it learns detection and elimination jointly in an end-to-end manner intending to jointly exploit the advantages of each other's improvements.

SID [3] designs depth networks to illuminate the shaded regions by estimating a linear transformation function. For penumbra (half-shadow), the shadow matting technique is used to handle it. The shadow regions are relit (lit), replacing the shaded areas of the shadow image. The fourth DSC [4] develops direction-aware attention mechanisms in spatial recurrent neural networks (RNNs) by introducing attention weights when aggregating spatial contextual features in the RNN. By training to learn these weights, we can recover the direction-aware spatial context (DSC) from detecting and eliminating shadows. The design is developed into a DSC module and embedded into a convolutional neural network (CNN) to learn different levels of DSC functions. RSI-GAN [5]designs a general framework for mining the information of residual and illumination through multiple GANs for shadow removal. DHAN [6] uses a dual-level aggregation network (DHAN) to eliminate boundary artefacts. DHAN has no downsampling and consists of growing inflated convolutions as a backbone for attention and prediction by aggregating multiple layers of contextual features. Auto-exposure [7] uses a fusion exposure approach, where a network with regular dithering of the exposure predicts the exposure parameters of the shadow part. A fusion network is used to learn the fusion parameters, the previous results are automatically fused to obtain a shadow-free image, and a refinement further removes the boundary residues. CANet [8] aims to mine the contextual information of the shadowed and non-shadowed regions. It is not just removal without detection; he relies on the global luminance averaging method to achieve the detection. The image from RGB to LAB only L channel is sensitive to shadows, the L here for global averaging, you can get a kind of fuzzy shadow-free map. You can distinguish between shadow-free and shadowed areas because the previous approaches of deep learning focus on increasing the perceptual field of the model and ignore the information of pairwise matching in the image. [31]–[34] are 2022-2023 years of self-supervised learning work for image shadow removal conclude diffusion model and Transformer.

B. Weakly supervised image shadow removal

Mask-ShadowGAN [9] argues that previous solutions to shadow removal problems using deep learning are supervised, paired Data. Using paired, there is no drawback for shadow removal methods. But the process of obtaining the paired dataset can be problematic. Therefore Mask-shadowGAN uses the idea of CycleGAN, introduce unpaired data to train the model, guiding the generation of shadow images by a shadow mask, solving the problem of multiple shadow maps corresponding to a shadow-free map, and modelling the relationship between shadow and shadow-free by the difference between shadow and shadow-free. LG- ShadowNet [10] argued that in practice, CNN training considers the ease of movement and unpaired data is more favoured for data collection. Therefore, they proposed a new Lightness-Guided Shadow Removal Network (LG-ShadowNet [10]) capable of training on unpaired data.

C. Unsupervised learning of image shadow removal

G2R [11] exploited the fact that shaded images usually contain both shaded and non-shaded regions. By this method, a set of shaded and unshaded patches can be cropped to construct unpaired data for network training, proposing three sub-network modules: shadow generation, shadow removal, and refinement. TC-GAN [16] performs the shadow removal task in an unsupervised manner. Comparing the GAN-based unsupervised shadow removal method with the bidirectional mapping in cyclic consistency, TC-GAN aims to learn a unidirectional mapping that converts the late-shadowed image into a shadow-free image. By the proposed goal consistency constraint aimed at connecting two GAN-based sub-networks, the correlation is between the shaded image and the output unshaded image, and the authenticity of the recovered unshaded image is strictly constrained.

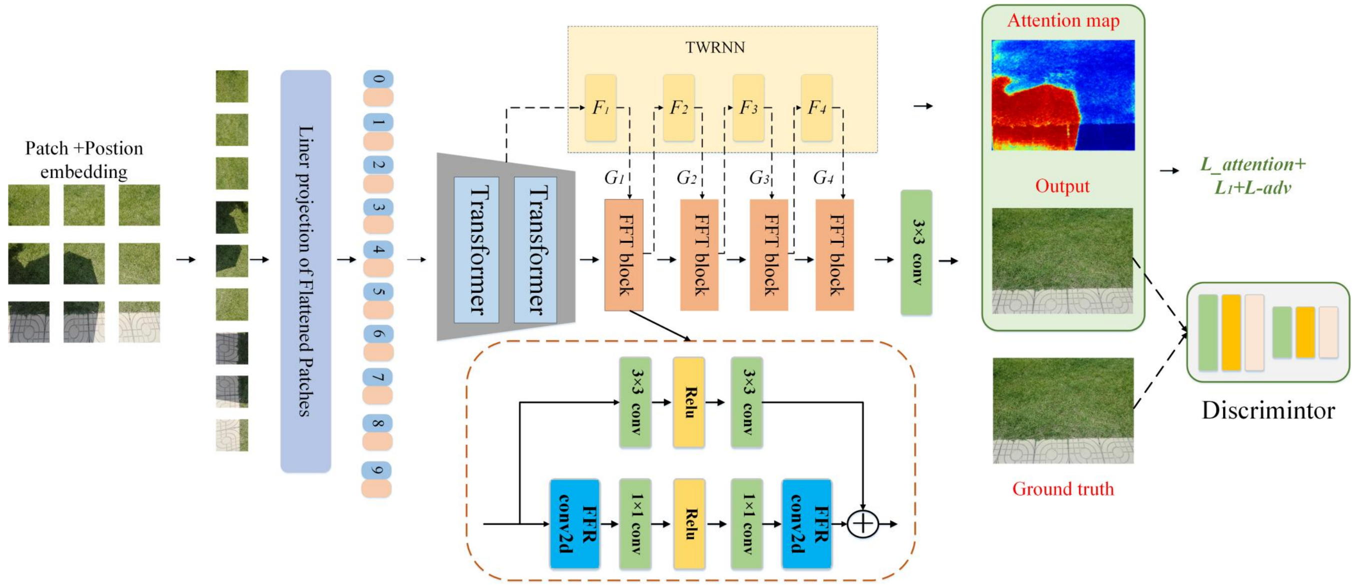

Fig. 2. The network structure in this paper, SpA-Former, is able to combine CNN-based local features with Transformer-based global representation to enhance the representation. spA-Former in general consists of Transformer encoding and CNN decoder coupled with GAN at the same time.

III. Methodology

The structure of SpA-Former is shown in Fig. 2. SpA-Former is inspired by [14], [15], [24], [27], [28], [30], [39] and divided into generators and discriminators. The generator consists of a Transformer encoder and a GAN network [36], [38]. First of all, the feature map is be patchEmbed by 32 patches, then into the Transformer encoder, following Fourier transform residuals(FTR) and Two-Wheel RNN Joint Spatial Attention(TWRNN), TWRNN is designed to make the network pay attention to specific shadow images, which can discover and find the focus map from the input element map. The attention graph is a two-dimensional matrix in which the value of each element is a constant value, indicating how much attention should be assigned to the pixel. The discriminator network consists of a series of convolution plus batch normalization and activation functions.

A. Vision Transformer Encoder and Gated Feed-Forward Network(GFFN)

The first thing to know is that this paper is not a traditional encoder-decoder structure as shown in Fig.2 and Fig. 3, SpA-Former only goes through the encoder of the transformer. Transformer branches are fed into the subsequent TWRNN(see section B) and FTR(see section C) to enhance the global perception capability of the CNN branch. And we must realize that the feature dimensions of CNN and Transformer are not the same. The dimensionality of CNN features is $C \times H \times W$ (C, H, and W are channel, height, and width, respectively), but the patch embedding dimension of Transformer is (K + 1)×E (K and E denote the number of image patches, respectively, we set E to 32 here because (TWRNN and FTR have a feature dimension of 32×480×480). When fedding from the Transformer branch to the TWRNN branch, there is a significant semantic gap between TWRNN features and Patch Embedding, because CNN feature mapping is obtained by local convolution operation, while Patch Embedding is aggregated by global self-attention mechanism. Therefore, FTR is applied to each block to gradually fill the semantic information gap.

*1) **Gated Feed-Forward Network(GFFN)**:* In the conventional encoder of a Transformer, the end layer is often output with MLP. The GFFN proposed in this paper consists of two fully connected layers with nonlinearity between the layers. This paper uses a gating mechanism to reformulate the linear transform layer to improve the flow of information through the network, as shown in Fig. 3, Transformer encoder contains the gating mechanism (implemented by a 1×1 convolution), which is formulated as a product of the components of two parallel-path linear transform layers. As in multi-self attention, deep convolution is included in the FFN to encode the positions of neighbouring pixels on the information space, helpful in learning local image structure for efficient recovery. Given an input tensor X,

$$\begin{aligned} \hat{\mathbf{X^3}} &= W_p^3 \text{ Gate } (\hat{\mathbf{X^2}}) \\ \text{Gate}(\mathbf{X^1}) &= \phi\left(W_d^1 W_p^1(\text{LN}(\mathbf{X}))\right) \odot W_d^1 W_p^1(\text{LN}(\mathbf{X})) \\ \text{Gate}(\mathbf{X^2}) &= \phi\left(W_d^2 W_p^2(\text{LN}(\mathbf{X}))\right) \odot W_d^2 W_p^2(\text{LN}(\mathbf{X})) \\ \text{Gate}(\hat{\mathbf{X^2}}) &= concat(\text{Gate}(\mathbf{X^1}), \text{Gate}(\mathbf{X^2})) \end{aligned} \tag{1}$$

where $\odot$ is the elemental multiplication, ϕ denotes the GELU nonlinear activation layer, and LN is the layer normalization. Overall, the GFFN controls the flow of information at each network level, allowing each level to focus on details that complement the other levels. In other words, when compared to multi self attention, GFFN plays a distinct role (focusing on enriching features with contextual information). Because the proposed GFFN performs more operations than

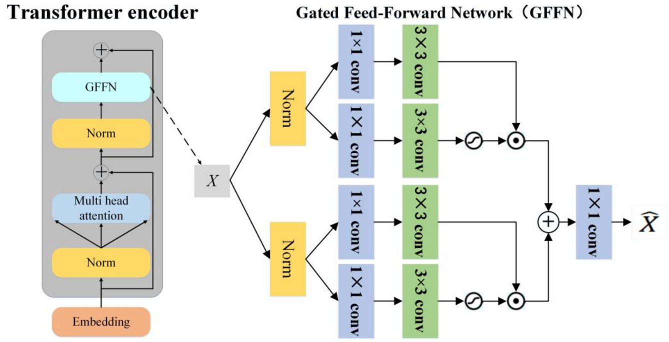

Fig. 3. Structure of Transformer encoder and Gated Feed-Forward Network(GFFN)

the conventional Vision Transformer, we reduce the expansion ratio, resulting in a parameter and computational burden that is comparable.

B. Two-Wheel RNN joint spatial Attention (TWRNN)

The Two-Wheel RNN Joint Spatial Attention model (TWRNN) is built based on the above two-round four-way RNN architecture. The RNN model is used to project the descent in four main directions. Another branch has been added to capture spatial context information and selectively highlight the expected shadow features. TWRNN can effectively identify areas affected by clouds according to the input shadow image.

Three standard residual blocks are first used to extract features to guide the three subsequent attention residual blocks to eliminate shadows by learning negative residual. Finally, the generated feature map is fed into two standard residual blocks to reconstruct the final shadow-removed image. In the following steps, the iterative update gradually focuses the attention on all shaded areas, and the shaded regions marked in red are marked more and more accurately.

F_1, F_2, F_3 and F_4 represent the eigenvalues in the four directions (up, down, left and right), while G_1, G_2, G_3 and G_4, represents the weight matrix corresponding to the four directions (up, down, left and right). TWRNN is a progressive generation of attentional maps, with a total of four graphs generated from light to deep depth, We have done the visualization results and ablation results for TWRNN in section IV-B1

C. Fourier transform residual block (FTR)

Deep Residual Fourier Transformation is introduced by [29], it is a common practice in end-to-end image recovery architectures to employ ResBlock, which learns the difference between shadow and clear images. Reconstruction of a non-shadowed image from a shaded counterpart requires changes to both low and high frequency information. Traditional ResBlock may be good at capturing the high-frequency components of an image, but it tends to ignore low-frequency information. Therefore, we design the Fourier transform residual block (FTR) as shown in Fig. 2, which combines the traditional resblock and Fourier convolution residual block.

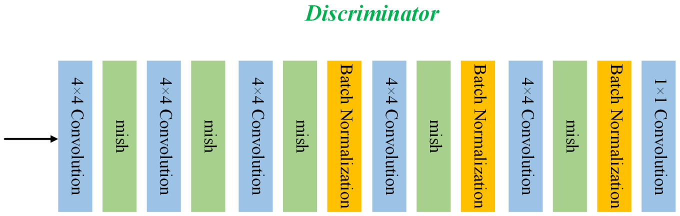

Fig. 5. The structure of discriminator

In addition, ResBlock is often unable to accurately model long-range information, which is important when reconstructing unshaded images from shadowed counterparts. The residual information is included to obtain the detected attention map by multiplication, and then it is used to obtain the negative residual, which is used to restore the image of the previous step to a slightly shaded or de-shadow image. We have done the visualization results and ablation results for FTR in section IV-B2

D. Discriminator

The structure of the discriminator is as Fig. 5, the input first goes through three 4×4 convolutions followed by Mish, then Batch normalization, then two more 4×4 convolutions followed by Mish and batch normalization, and finally 1×1 convolution.

E. Loss function

The loss function is mainly composed of five items: L_{CGAN}, L_1, and $L_{attention}$.

$$L_{CGAN}(G,D) = \mathcal{E}_{x,y\sim p_{dda}(x,y)}[\log D(x,y)]+ \mathcal{E}_{x\sim p_{data}(x),z\sim p_z(z)}[\log(1-D(x,G(x,z)))] \tag{2}$$

The second part of the loss is the standard L_1 loss, which is used to measure the accuracy of each reconstructed pixel. As shown in (2), I_{input} and I_{output} are the input and output images, respectively, λ_C is the weight of each channel, ϕ is the predicted result of the network, and C, H and W indicate the number of channels, the height and the width of the image, respectively.

$$L_1(G) = \frac{1}{4HW}\sum_{c=1}^{C}\sum_{v=1}^{H}\sum_{u=1}^{W}\lambda_c\left|I_{\text{outpud}}^{(u,v,c)} - \phi\left(I_{\text{input}}\right)^{(u,v,c)}\right|_1 \tag{3}$$

The third part of the loss is attention loss, which is defined as (6). The matrix A is the attention map module generated by soft attention, and matrix M is the binary image of the shadow area, which is composed of shadow and no shadow images.

$$L_{Attention} = \|A-M\|_2^2 \tag{4}$$

The total loss function is:

$$L_{total} = L_1 + L_{CGAN} + L_{Attention} \tag{5}$$

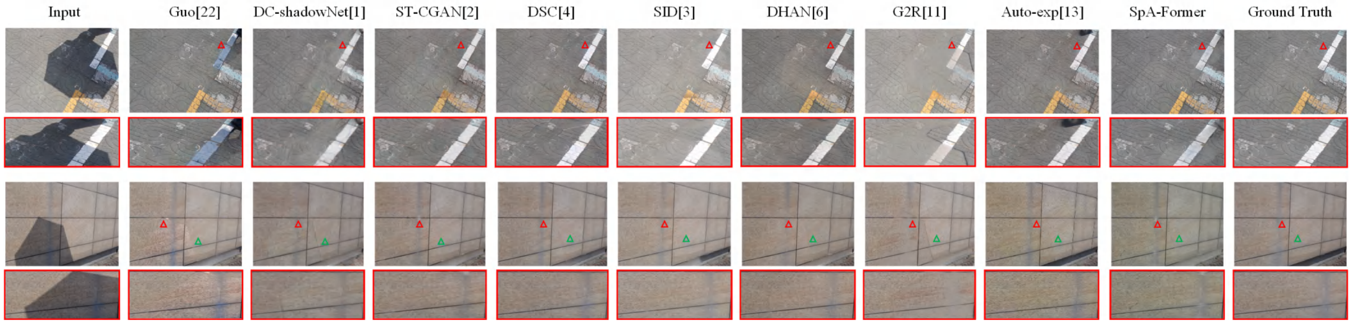

Fig. 4. The comparative results of SpA-Former with other exsiting methods

IV. Experiment

A. Dataset and performance comparisons with existing methods

This paper uses the latter dataset ISTD [2], SRD. We chose the Adam optimizer with a β of 0.999 at the same time.

Our method is compared with existing methods including Yang [26], Guo [23], Gong [25], DeShadowNet [1], STC-GAN [2], DSC [4], Mask-ShadowGAN [9], RIS-GAN [5], DHAN [6], SID [3], LG-shadow [10], G2R [11], SG-ShadNet [12], Auto-exp [13]. We adopt RMSE, SSIM and PSNR in the LAB color space as evaluation metrics.

We evaluate the performance of different methods on the shadow regions, non-shadow regions, and the whole image. We can see that the proposed SpA-Former achieves the satisfactory performance among all the compared methods as shown in Fig.4. We also report the shadow removal performance of our proposed method on the ISTD and srd dataset. As shown in Table.I. According to the Table, our suggested SpA-Former gets the approving RMSE values in shadow areas, non-shadow regions, and the complete image on the ISTD. This implies that the recovered shade-removal pictures generated by our SpA-Former are substantially closer to the corresponding ground-truth shadow-free images.

B. Ablation Study on SpA-Former

*1) **Quantitative comparisons in Two-wheel RNN joint spatial attention(TWRNN)**:* To validate the effectiveness of Two-wheel RNN joint spatial attention, we make the ablation experiment in Fig.6. Semi-supervised learning is utilized to train the model in this work since SpA-Former does not require shaded mask image, semi-supervised learning is used to demonstrate that the model can fully utilize shaded images to increase model performance. To further illustrate the RNN joint spatial attention model, we can see from the figure that TWRNN produces a total of four attention maps, from attention1 to attention4, the heat map of shadows is gradually obvious and the difference between shaded and non-shaded regions is well represented. In this paper, after choosing TWRNN, SpA-Former can effectively handle the color of the whole image information and generate a more realistic attention image with shadow detection.

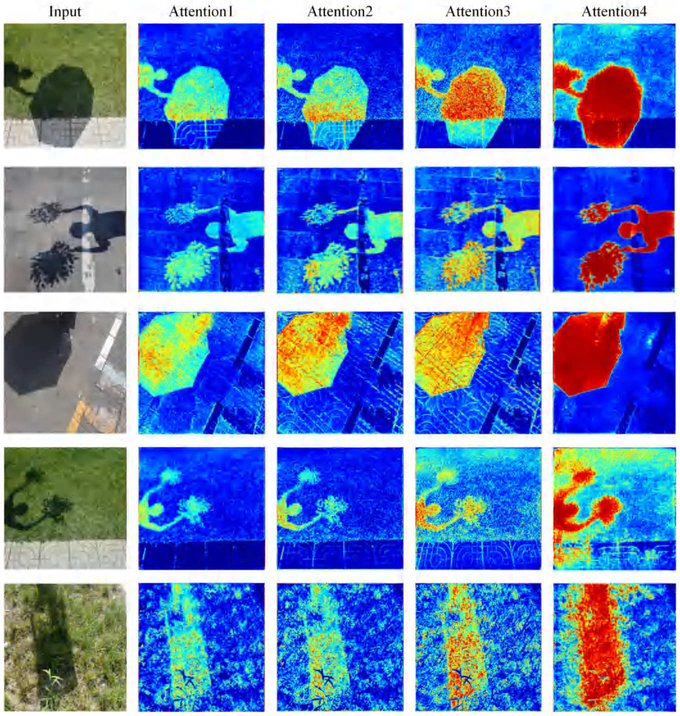

Fig. 6. Ablation of TWRNN, the first line is the input, attention1 to attention4 is the progressive attention graph generation of TWRNN

*2) **Quantitative comparisons in Transformer encoder and Fourier Transform Residuals Module(FTR)**:* As shown in III, by added Transformer encoder or Two-wheel RNN(TWRNN)/Fourier Transform Residuals Module(FTR) could get a significant improvement over classical encoder-decoder. And the combination of them could achieve the best performance, indicating the regulation is beneficial for the task. The combination of these proposed three parts could recover the details and generate realistic results in easy nature and complicated ISTD dataset. The addition of the Transformer encoder effectively increases the PSNR and ssm to 0.84 and 26.61, respectively, clearly enhancing the significant texture features. After adding the FTR layer, the PSNR rises to 26.87,ssim rises to 0.92, showing the superiority of this module in shadow removal. With the addition of TWRNN, PSNR rises to 26.82 and ssim rises to 0.91. After all three modules are added,

TABLE I
PERFORMANCE COMPARISON OF SHADOW REMOVAL ON ISTD (RED MEANS RANKING FIRST, BLUE MEANS RANKING SECOND)

Models	Venue	RMSE	RMSE-N	RMSE-S	SSIM	SSIM-N	SSIM-S	PSNR	PSNR-N	PSNR-S	**Paramas**	**Flops**
Traditional methods												
Yang [26]	TIP2012	15.63	14.83	19.82	-	-	-	-	-	-	-	-
Guo [23]	TPAMI2013	9.3	7.46	18.95	0.919	0.944	0.978	23.07	24.86	30.98	-	-
Gong [25]	BMVC2014	8.53	7.29	14.98	0.908	0.929	0.98	24.07	25.26	32.43	-	-
Supervised learning methods												
DeShadowNet [1]	CVPR2017	7.83	7.19	12.76	-	-	-	-	-	-	-	-
STC-GAN [2]	CVPR2018	7.47	6.93	10.33	0.929	0.947	0.985	27.43	28.67	35.8	29.24M	17.88G
DSC [4]	TPAMI2018	6.67	6.39	9.22	0.845	0.885	0.967	26.62	28.18	33.45	22.30M	123.47G
SG-ShadowNet [12]	ECCV2022	-	-	-	-	-	-	-	-	-	6.2M	39.7G
Half-Supervised learning methods												
Mask-ShadowGAN [9]	ICCV2019	7.61	7.03	10.35	-	-	-	-	-	-	11.38M	56.83G
RIS-GAN [5]	AAAI2019	6.62	6.31	9.15	-	-	-	-	-	-	-	-
DHAN [6]	CVPR2019	6.28	5.92	8.43	0.921	0.941	0.983	27.71	29.54	34.79	21.75M	262.87G
SID [3]	ICCV2019	7.96	7.72	9.64	0.948	0.964	0.986	25.01	26.1	32.88		
LG-shadow [10]	ECCV2020	6.67	5.93	11.51	0.906	0.938	0.974	25.83	28.32	31.08	141.2M	39.8G
Un-Supervised methods												
G2R [11]]	CVPR2021	7.84	7.54	10.71	0.931	0.967	0.974	24.72	26.18	31.62	22.8M	113.9G
Auto-Exp] [13]	CVPR2021	5.88	5.51	7.9	0.845	0.879	0.975	27.19	28.6	34.71	142.2M	104.8G
SpA-Former		5.06	6.66	5.89	0.931	0.956	0.982	27.73	30.16	34.71	0.47MB	15G

TABLE II
PERFORMANCE COMPARISON OF IMAGE SHADOW REMOVAL ON SRD. (THE RED MARKED REPRESENTATIVE RANKED FIRST AND THE BLUE MARKED REPRESENTATIVE RANKED SECOND

Models	Venue	RMSE	RMSE-N	RMSE-S	SSIM	SSIM-N	SSIM-S	PSNR	PSNR-N	PSNR-S
Traditional methods										
Guo [23]	TPAMI2013	14.05	6.47	29.89	-	-	-	-	-	-
Supervised learning methods										
DeShadowNet [1]	CVPR2017	6.64	4.84	11.78	-	-	-	-	-	-
DSC [4]	TPAMI2018	17.43	16.45	20.09	0.648	0.760	0.905	20.65	23.35	25.83
Half-Supervised learning methods										
DHAN [6]	CVPR2019	4.80	3.77	6.76	0.928	0.967	0.972	28.93	32.49	32.63
Un-Supervised methods										
Auto-exp [11]]	CVPR2021	6.75	5.97	9.02	0.856	0.920	0.956	26.72	29.37	31.43
SpA-Former	-	6.60	5.71	8.94	0.920	0.964	0.966	27.49	31.11	31.10

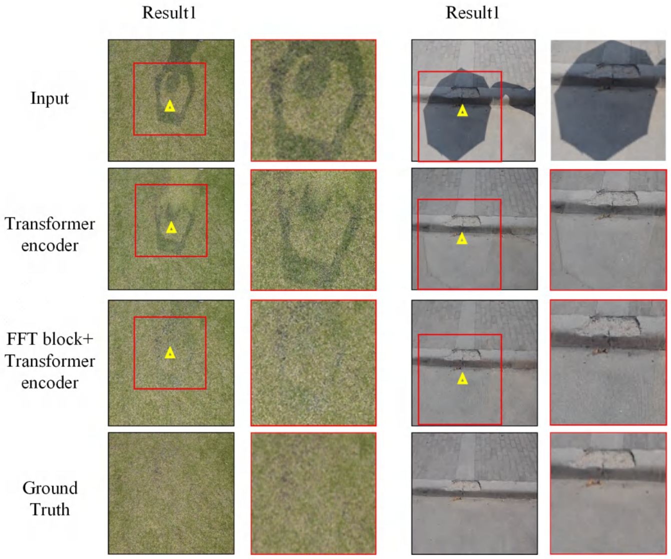

Fig. 7. Ablation of SpA-Former (The first line is the baseline, the second line is the result of adding the Transformer layer, the third line is the result of the Fourier transform residual block and the Transformer layer, and the fourth line is the Ground Truth)

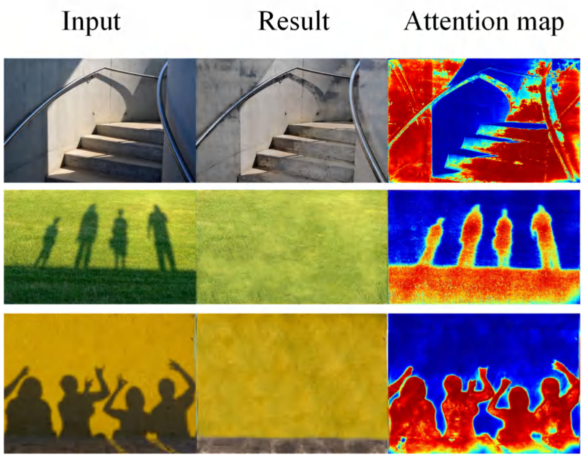

Fig. 8. Results of SpA-Former on real world image(Three scenes, respectively, stairs, grass, playground)

TABLE III
ABLATION STUDY OF SPA-FORMER ON THE ISTD DATASET.

Transformer encoder	FTR	TWRNN	RMSE	SSIM	PSNR
√			7.87	0.84	26.21
	√		6.56	0.92	26.87
		√	6.91	0.91	26.82
√	√	√	**5.06**	**0.931**	**27.73**

PSNR and ssim reach a maximum of 27.73 and 0.931.

C. Generalization experiments

We chose three real-world scenes: stairs, playground, and square, and the SpA-Former shadow removal results in these three scenes are shown in Fig. 8. The results on the green and square are relatively good, but shadow removal on the stairs can be further improved. Considering that ISTD/SRD do not include staircases in this manner, they can be viewed as self-supervised learning that extends beyond the scope of the network. At the same time, this provides us with some future directions. We can further optimize and apply the network to a wide range of complex scenes rather than just a single playground, meadow, or square.

V. Discussion of SpA-Former

It is very important to note that SpA-Former is not a simple stack, but a simple and lightweight network derived from rigorous theory and experiments. The Params and Flops of SpA-Former are made to be the smallest so far. Although the proposed SpA-Former in this paper doesn't achieve the best perform in terms of effect and metrics, its contribution is that we design a lightweight and deployable algorithm. Currently, %90 image de-shadowing algorithms are two-stage, and most of them are even ternary data, which need black and white mask maps of shadows to in the training phase. In contrast, we only need binary data (real and shaded maps), and we reach the top 3 levels in terms of metrics. Our Params is currently the smallest of among all methods, reaching only 0.47MB, and Flops reaches 15G, achieving Lightweight. However, the results of SpA-Former will invariably be inferior to the previous algorithm in achieving lightness. Simultaneously, this paper has a lot of room for improvement; for example, in Fig. 6 and Fig. 8, some attention maps for grass or complex shadow scenes cannot be 100% accurate, so our future direction could be to design the network very precisely so that it can play a good deployment and assistance in the unmanned, or robot vision.

VI. Results

This paper proposes SpA-Former, a one-stage lightweight shadow removal network. The SpA-Former integrates detection and removal into a one-stage network that learns the mapping function between shadows and no shadows, eliminating the need for a separate shadow detection stage or post-processing adjustments. SpA-Former is currently the lightest image de-shadowing algorithm available, with a parameter count of 0.47 MB and an inference speed of 0.0459 s. SpA-former can adapt to the de-shadowing of different semantic regions of shadows in the real world, according to experimental results on two datasets and natural scenes.

VII. Acknowledgment

This work is supported by National Natural Science Foundation 62273235, Chinese Ministry of Education Research Found No. 8091B022101.

References

[1] Qu L, Tian J, He S, et al. Deshadownet: A multi-context embedding deep network for shadow removal//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 4067-4075.

[2] Wang J, Li X, Yang J. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1788-1797.

[3] Le H, Samaras D. Shadow removal via shadow image decomposition//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 8578-8587.

[4] Hu X, Fu C W, Zhu L, et al. Direction-aware spatial context features for shadow detection and removal. IEEE transactions on pattern analysis and machine intelligence, 2019, 42(11): 2795-2808.

[5] Zhang L, Long C, Zhang X, et al. Ris-gan: Explore residual and illumination with generative adversarial networks for shadow removal//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 12829-12836.

[6] Cun X, Pun C M, Shi C. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting GAN//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 10680-10687.

[7] Fu L, Zhou C, Guo Q, et al. Auto-exposure fusion for single-image shadow removal//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 10571-10580.

[8] Chen Z, Long C, Zhang L, et al. CANet: A Context-Aware Network for Shadow Removal//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 4743-4752.

[9] Hu X, Jiang Y, Fu C W, et al. Mask-ShadowGAN: Learning to remove shadows from unpaired data//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 2472-2481.

[10] Liu Z, Yin H, Mi Y, et al. Shadow removal by a lightness-guided network with training on unpaired data. IEEE Transactions on Image Processing, 2021, 30: 1853-1865.

[11] Vasluianu F A, Romero A, Van Gool L, et al. Self-Supervised Shadow Removal. arXiv preprint arXiv:2010.11619, 2020.

[12] Wan J, Yin H, Wu Z, et al. Style-Guided Shadow Removal//European Conference on Computer Vision. Springer, Cham, 2022: 361-378.

[13] Fu L, Zhou C, Guo Q, et al. Auto-exposure fusion for single-image shadow removal//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 10571-10580.

[14] Mei X, Ye X, Zhang X, et al. UIR-Net: A Simple and Effective Baseline for Underwater Image Restoration and Enhancement. Remote Sensing, 2022, 15(1): 39

[15] Qiu L, Yu D, Zhang C, et al. A LocalCGlobal Framework for Semantic Segmentation of Multisource Remote Sensing Images[J]. Remote Sensing, 2022, 15(1): 231.

[16] Le H, Samaras D. From shadow segmentation to shadow removal//European Conference on Computer Vision. Springer, Cham, 2020: 264-281.

[17] Tan C, Feng X. Unsupervised Shadow Removal Using Target Consistency Generative Adversarial Network. arXiv preprint arXiv:2010.01291, 2020.

[18] Liu Z, Yin H, Wu X, et al. From shadow generation to shadow removal//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 4927-4936.

[19] Tan C, Feng X. Unsupervised Shadow Removal Using Target Consistency Generative Adversarial Network. arXiv preprint arXiv:2010.01291, 2020.

[20] He Y, Xing Y, Zhang T, et al. Unsupervised Portrait Shadow Removal via Generative Priors//Proceedings of the 29th ACM International Conference on Multimedia. 2021: 236-244.

[21] Zhu Y, Xiao Z, Fang Y, et al. Efficient Model-Driven Network for Shadow Removal. 2022.

[22] Zhu Y, Huang J, Fu X, et al. Bijective Mapping Network for Shadow Removal//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 5627-5636.

[23] R. Guo, Q. Dai, and D. Hoiem, Paired regions for shadow detection and removal, IEEE TPAMI, vol. 35, no. 12, 2012.

[24] Zhang X, Chen F, Wang C, et al. Sienet: Siamese expansion network for image extrapolation. IEEE Signal Processing Letters, 2020, 27: 1590-1594.

[25] H. Gong and D. Cosker, Interactive shadow removal and ground truth for variable scene categories, in Proc. BMVC, 2014.

[26] Q. Yang, K. Tan, and N. Ahuja. Shadow removal using bilateral filtering. IEEE TIP, 21(10):4361??C4368, 2012.

[27] Pan H. Cloud removal for remote sensing imagery via spatial attention generative adversarial network. arXiv preprint arXiv:2009.13015, 2020.

[28] Zamir S W, Arora A, Khan S, et al. Restormer: Efficient transformer for high-resolution image restoration//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 5728-5739.

[29] Mao X, Liu Y, Shen W, et al. Deep residual fourier transformation for single image deblurring. arXiv preprint arXiv:2111.11745, 2021.

[30] Shen R, Zhang X, Xiang Y. AFFNet: attention mechanism network based on fusion feature for image cloud removal. International Journal of Pattern Recognition and Artificial Intelligence, 2022: 2254014.

[31] Yu Q, Zheng N, Huang J, et al. CNSNet: A Cleanness-Navigated-Shadow Network for Shadow Removal. arXiv preprint arXiv:2209.02174, 2022.

[32] Jin Y, Yang W, Ye W, et al. ShadowDiffusion: Diffusion-based Shadow Removal using Classifier-driven Attention and Structure Preservation[J]. arXiv preprint arXiv:2211.08089, 2022.

[33] Wan J, Yin H, Wu Z, et al. CRFormer: A Cross-Region Transformer for Shadow Removal. arXiv preprint arXiv:2207.01600, 2022.

[34] Xu Y, Lin M, Yang H, et al. Shadow-Aware Dynamic Convolution for Shadow Removal. arXiv preprint arXiv:2205.04908, 2022.

[35] Zhu T, Xia S, Bian Z, Lu C. Highlight removal in facial images. InPattern Recognition and Computer Vision: Third Chinese Conference, PRCV 2020, Nanjing, China, October 16C18, 2020, Proceedings, Part I 3 2020 (pp. 422-433). Springer International Publishing.

[36] Wu X, Lu C, Gu C, Wu K, Zhu S. Domain adaptation for viewpoint estimation with image generation. In2021 International Conference on control, automation and information sciences (ICCAIS) 2021 Oct 14 (pp. 341-346). IEEE.

[37] Lu C, Wang H, Gu C, Wu K, Guan X. Viewpoint estimation for workpieces with deep transfer learning from cold to hot. InNeural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13-16, 2018, Proceedings, Part I 25 2018 (pp. 21-32). Springer International Publishing.

[38] Lu C, Gu C, Wu K, Xia S, Wang H, Guan X. Deep transfer neural network using hybrid representations of domain discrepancy. Neurocomputing. 2020 Oct 7;409:60-73.

[39] Lu C, Zhu H, Koniusz P. From Saliency to DINO: Saliency-guided Vision Transformer for Few-shot Keypoint Detection. arXiv preprint arXiv:2304.03140. 2023 Apr 6.

[40] Jiang Y, Yang F, Bian Z, Lu C, Xia S. Mask removal: Face inpainting via attributes. Multimedia Tools and Applications. 2022 Sep;81(21):29785-97.

[41] Lu C, Xia S, Shao M, Fu Y. Arc-support line segments revisited: An efficient high-quality ellipse detection. IEEE Transactions on Image Processing. 2019 Aug 15;29:768-81.

[42] Lu C, Xia S, Huang W, Shao M, Fu Y. Circle detection by arc-support line segments. In2017 IEEE International Conference on Image Processing (ICIP) 2017 Sep 17 (pp. 76-80). IEEE.